TUNIS BUSINESS SCHOOL

WEB SCRAPPER

NewsFuse

Summary

Overview | Main concepts | High-Level Design | Tools | Code & visualization





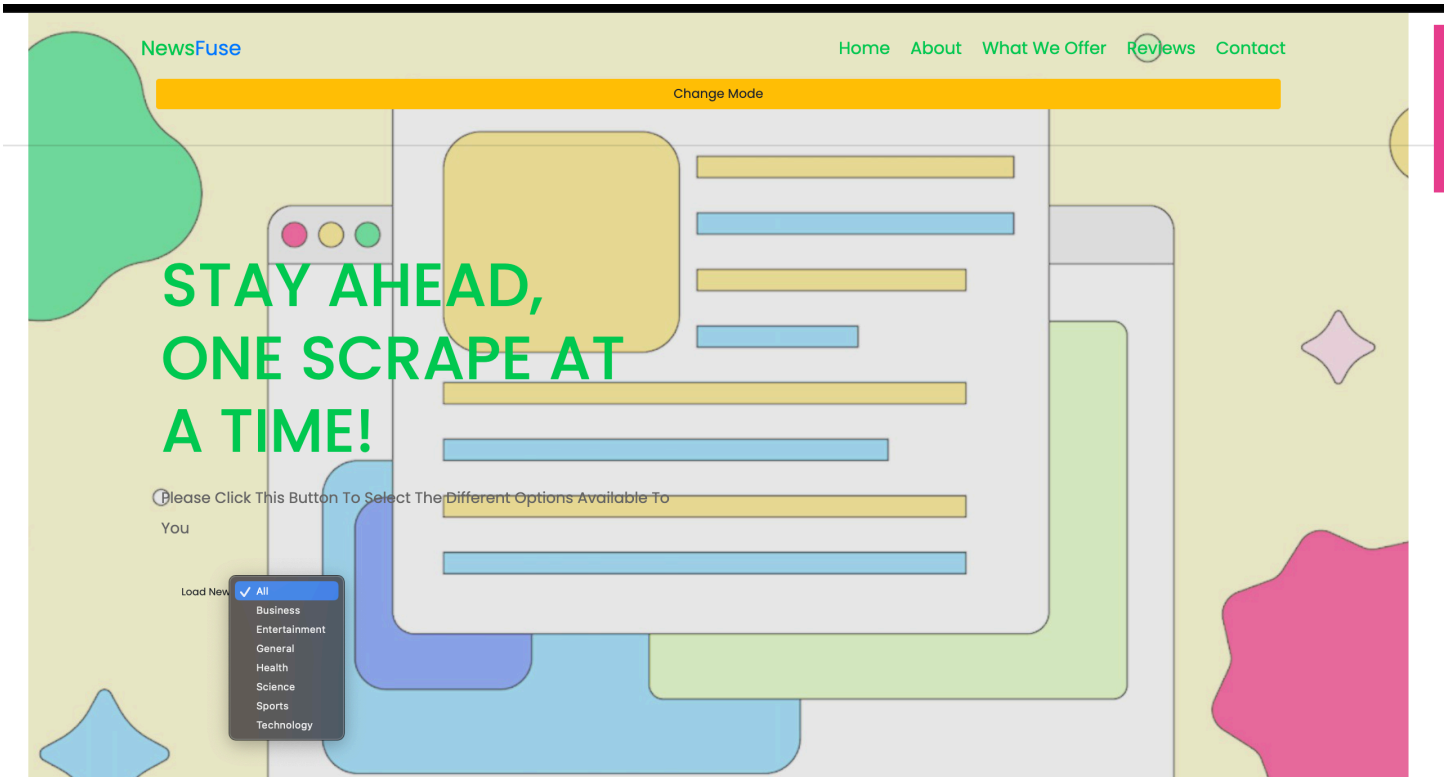
Yosr Jaouadi

Zeineb Belaid

Overview

NewsFuse is an innovative web scraper and news aggregation tool designed to collect, process, and present news articles providing users with a centralized platform to access the latest news grouped into various categories.

What We Achieved



Website interface (Safari)

Articles display (Chrome)

Main Concepts

Overview Components:

- Web Scraping

Functional Flow:

Data Collection

- Data Aggregation
- News Categorization

Data Processing
Data Presentation

Pros & cons

Advantages

- **Customizability**: Tailored scraping to specific needs.
- **Freshness**: Real-time updates.
- **Ease of Use**: user-friendly, with intuitive interfaces and straightforward setup processes.
- **Cost-Efficiency**: Lower cost compared to subscription-based services.

Limitations

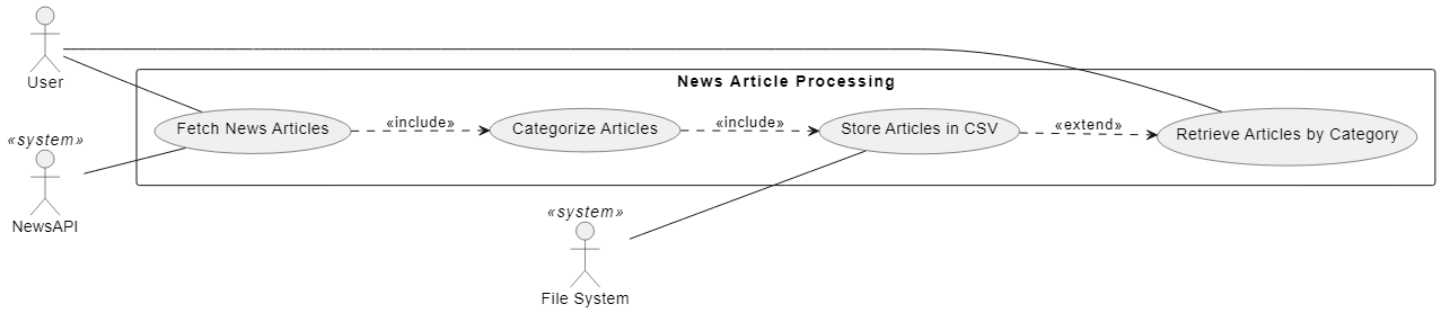
- **Scalability**: Challenges in handling large-scale data.
- **Maintenance**: Regular updates are needed for scraping scripts.
- **Legal Issues**: Compliance with web scraping policies.
- **Layout Instability**: Websites frequently update their layout and structure, which can cause web scraping tools to fail if not regularly monitored and updated.

High-level Design

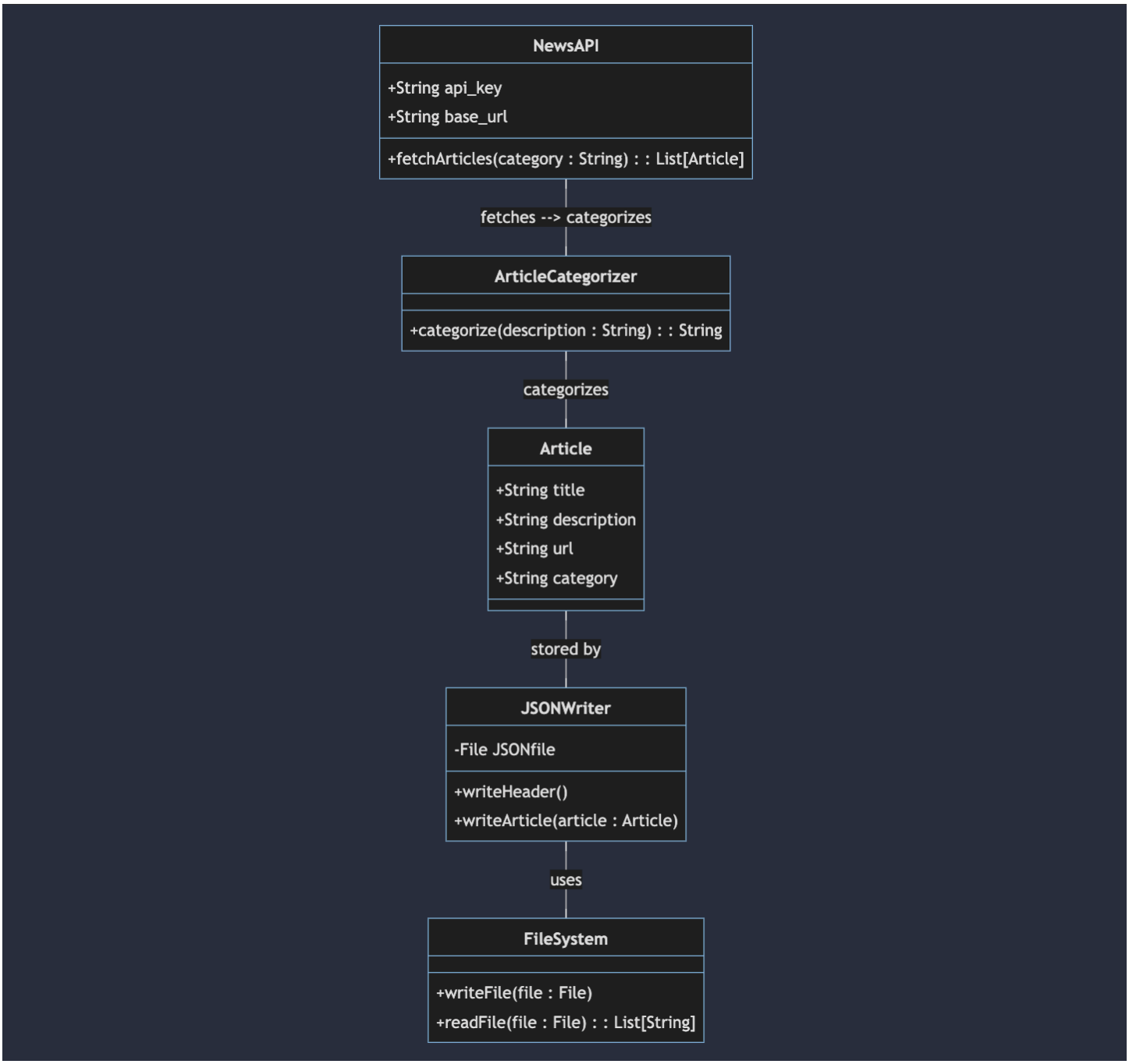
Inputs & Outputs:

- Input: URLs, Keywords, Time Intervals
- Output: Aggregated News Data, categorized articles

Use Case Diagram:



Class Diagram:

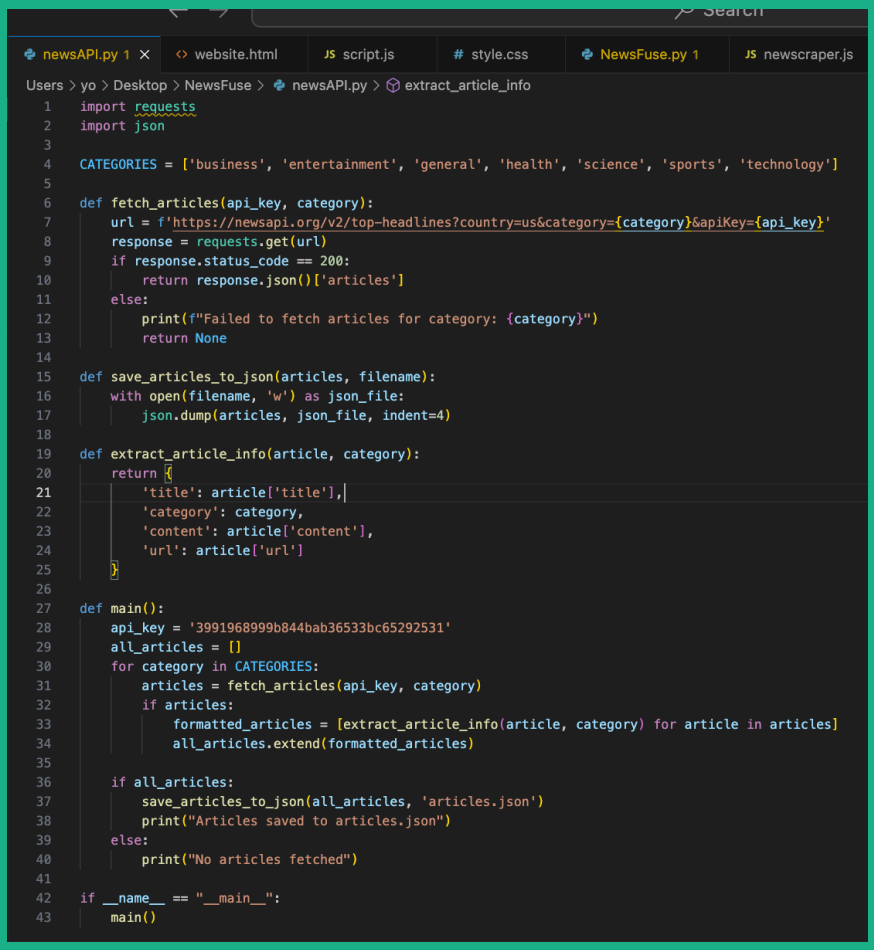


Tools		Development phases	

	• **Task**: Web Scraping
	  ▪ **Tools**: Python
	• **Task**: Data Storage
	  ▪ **Tools**: JSON file
	• **Task**: Front-end Development
	  ▪ **Tools**: HTML, CSS, JavaScript
	  ▪ **IDE**: Visual Studio Code
			1. Requirement Analysis
			Define project goals, functionalities, and user requirements.
			2. Design
			Illustrates system architecture and components.
			3. Implementation
			Develop the web scraping engine to extract data from newsAPI website then store articles and display them according to the user preferences.
			4. Testing
			executes test cases to ensure the functionality, performance, and security of the web scrapping.
			5. Deployment
			Provide ongoing maintenance, updates, and support to ensure our NewsAggregator remains functional and up-to-date.
			Also,address any user feedback or issues reported after deployment.

Code

The solution will consist of a web scraping tool built using Python, utilizing the ***requests*** library to scrape news articles from ***newsAPI*** website.



Python

Executable Code

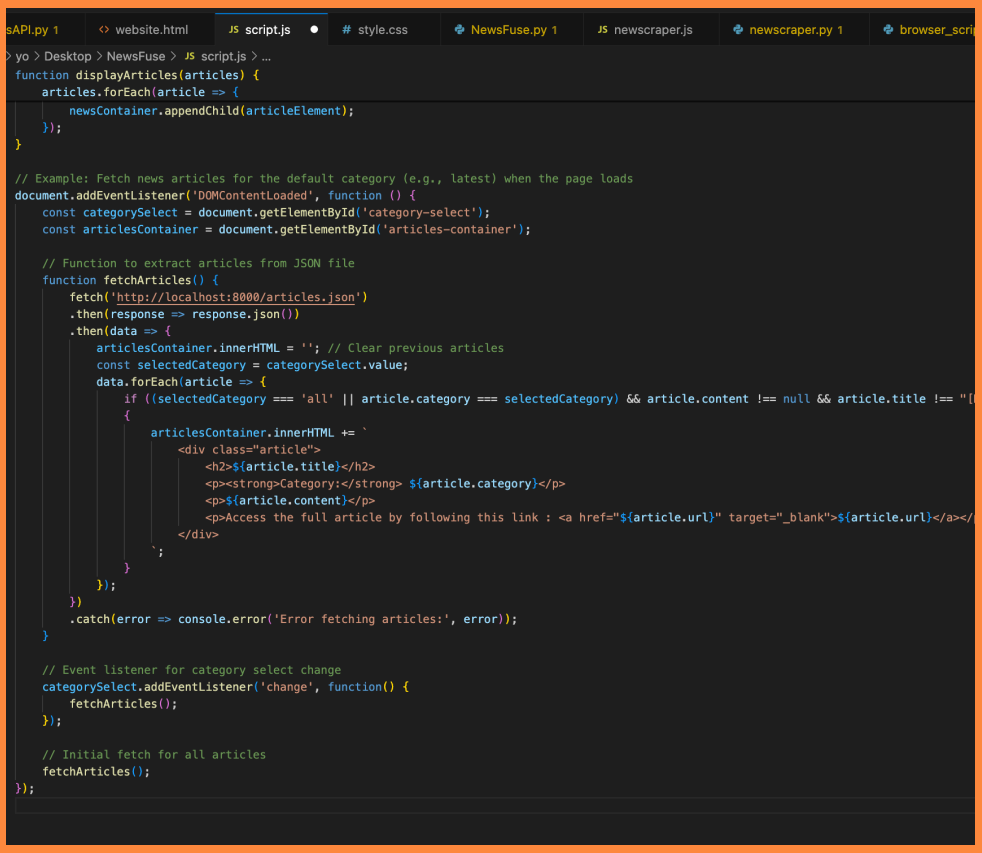
Articles are categorized into predefined topics: Business, Entertainment, Health, Science, Sports, Technology, and General.

This HTML structure forms the skeleton of the NewsFuse web page.

It includes a container (articles_container) where the news articles will be displayed and links to the CSS stylesheet (styles.css) and JavaScript file (script.js).

	HTML

Provides users with a web interface to browse articles by topic and access specific articles’ categories.



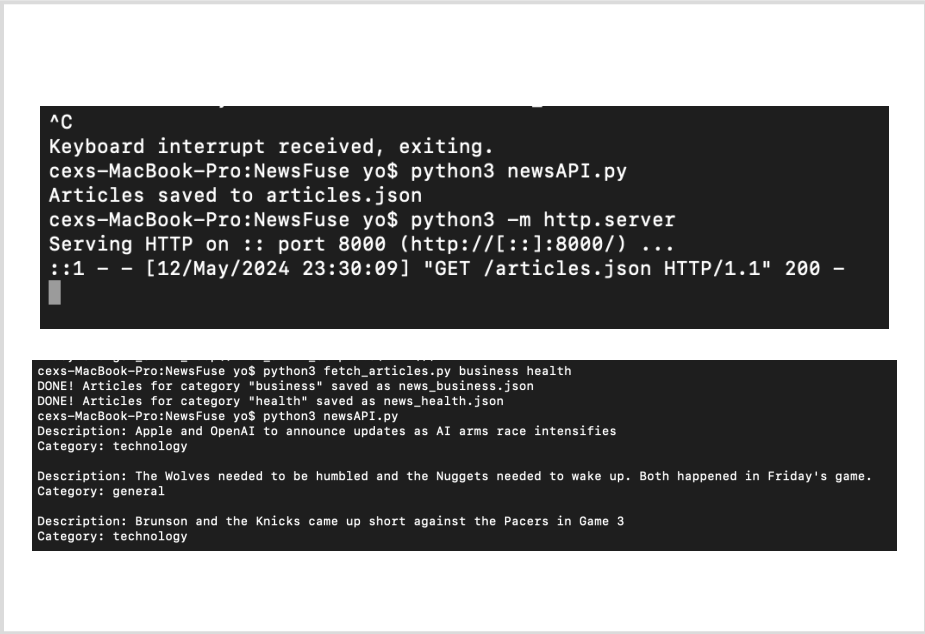
JavaScript

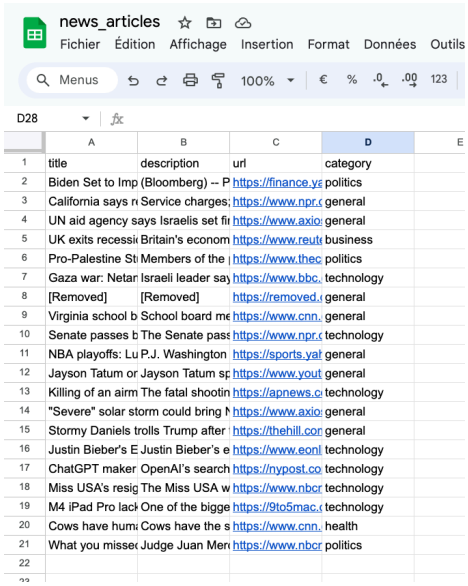
This JavaScript snippet fetches the news data from the Desktop/*NewsFuse/articles.json* endpoint, parses the JSON response, and dynamically creates and appends HTML elements to display the news articles on the webpage.

This is it ! A well-designed web scraper for a news aggregator combines technical expertise with ethical considerations. By understanding the theoretical aspects and implementing best practices, we created a reliable and efficient system for collecting news data.

Code Output

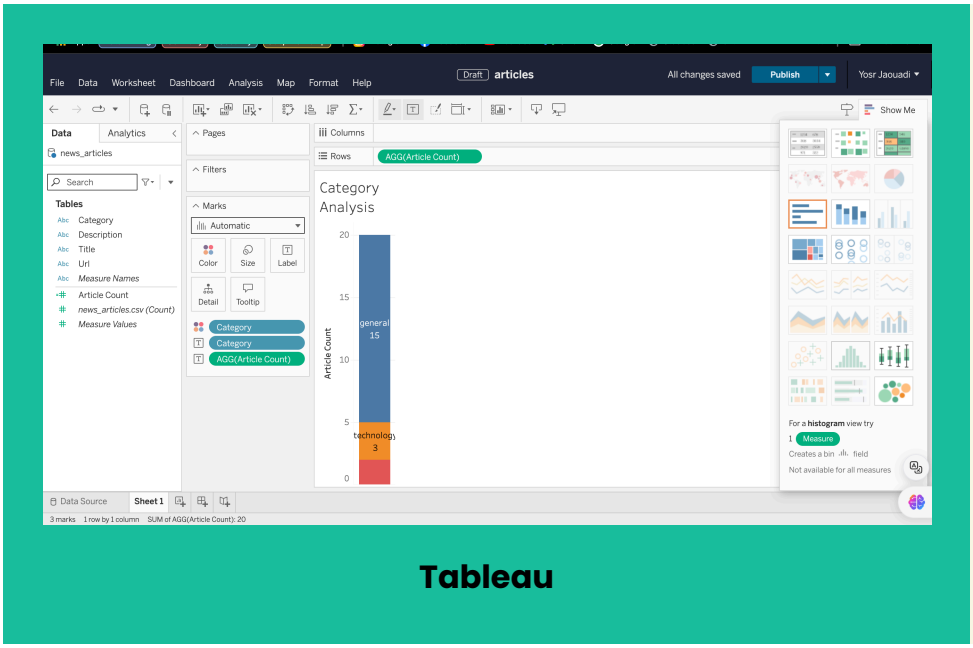
A JSON file will be used each time to store the scraped data, displaying articles organized by title, description, URL, and category.





Data Visualization

To provide a preliminary insight into the capabilities of NewsFuse, we conducted a trial visualization using ***Tableau***.



Tableau

1.We connected to our dataset stored in a CSV file.

-which includes news articles categorized under general news, weather, and technology.

2. we created a **category distribution visualization** to analyze the proportion of news articles within these categories presented as a stacked bar chart.

This trial demonstrates just a snippet of what category analysis can offer, highlighting how ***Tableau*** can effectively be used to uncover trends and patterns in our aggregated news data.

Potential Enhancements and Future Work

- [ ] **Advanced Data Visualization**:

Implement graphs and charts to show news trends, popular topics, and article distribution using tools like Tableau.

- [ ] **User Analytics**:

Analyze user interaction data to understand visitor demographics, preferences, and behavior.

- [ ] **Personalization**:

Implement machine learning algorithms to personalize news feeds based on user interests and reading history.

- [ ] **Content Analysis**:

Perform sentiment analysis on news articles to provide insights into the general tone of the news for handling content takedown requests and user data privacy.

Contact Us

For further information, please reach out to us

Emails: jaouadi.yosr@tbs.u-tunis.tn **&** zaineb6belaid@gmail.com

Github:

Yoyo33333333/
NewsFuse

System Assurance and Security Project

0 Contributors 0 Issues 0 Stars 0 Forks